Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Theoretical and empirical properties of model assisted decision-based regression estimators

by Jun Shao, Eric Slud, Yang Cheng, Sheng Wang
and Carma Hogue

June 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, June 2014
Vol. 40, No. 1, pp. 81-104
Statistics Canada, Catalogue No. 12-001-X

Theoretical and empirical properties of model assisted decision-based regression estimators

Jun Shao, Eric Slud, Yang Cheng, Sheng Wang, and Carma Hogue[1]

Abstract

In 2009, two major surveys in the Governments Division of the U.S. Census Bureau were redesigned to reduce sample size, save resources, and improve the precision of the estimates (Cheng, Corcoran, Barth and Hogue 2009). The new design divides each of the traditional state by government-type strata with sufficiently many units into two sub-strata according to each governmental unit's total payroll, in order to sample less from the sub-stratum with small size units. The model-assisted approach is adopted in estimating population totals. Regression estimators using auxiliary variables are obtained either within each created sub-stratum or within the original stratum by collapsing two sub-strata. A decision-based method was proposed in Cheng, Slud and Hogue (2010), applying a hypothesis test to decide which regression estimator is used within each original stratum. Consistency and asymptotic normality of these model-assisted estimators are established here, under a design-based or model-assisted asymptotic framework. Our asymptotic results also suggest two types of consistent variance estimators, one obtained by substituting unknown quantities in the asymptotic variances and the other by applying the bootstrap. The performance of all the estimators of totals and of their variance estimators are examined in some empirical studies. The U.S. Annual Survey of Public Employment and Payroll (ASPEP) is used to motivate and illustrate our study.

Key Words: Asymptotic normality; Bootstrap; Decision-based estimator; Probability proportional to size; Stratification; Variance estimation.

1 Introduction

The U.S. Annual Survey of Public Employment and Payroll (ASPEP) provides current estimates for full- and part-time state and local government employment and payroll classified by government functions (such as: elementary and secondary education, higher education, police protection, fire protection, financial administration, judicial and legal, *etc.*). This survey covers state and local government units (89,526 according to the 2007 Census of Governments), which include counties, cities, townships, units called "special districts", and school districts. ASPEP is the only source of public employment data by government function and job category, providing data on numbers of full- and part-time employees and payroll, as well as on hours worked by part-time employees. Data collection usually begins in March and continues for about seven months, with the pay period containing March 12 as reference period.

Let U denote the finite population of N units stratified into H strata, $U_1,\ldots,U_H$, where U_h contains N_h units and $N_1 + \cdots + N_H = N$. The traditional sampling design for the ASPEP is a stratified probability proportional to size (PPS) design, where the strata are constructed using state and the government types, which are county, subcounty (city or town), special district, or school district. The size of each unit is the total payroll, and sampling across strata is independent. In 2009, a modified sampling design was developed, which cuts some strata U_h into two sub-strata, U_{h1} and U_{h2} with N_{h1} and N_{h2} units, respectively, where U_{h1} contains smaller-size units (Cheng *et al.* 2009). The idea was to save

1. Jun Shao, Statistics Department University of Wisconsin, Madison WI , E-mail: shao@stat.wisc.edu; Eric Slud, Center for Statistical Research and Methodology, US Census Bureau, Washington DC and Mathematics Department, University of Maryland, College Park, MD, E-mail: eric.v.slud@census.gov; Yang Cheng, Demographic Statistical Methods Division, US Census Bureau, Washington DC, E-mail: yang.cheng@census.gov ; Sheng Wang, Mathematica Policy Research, Princeton NJ, E-mail: swang@mathematica-mpr.com; and Carma Hogue, Governments Division, US Census Bureau, Washington DC, E-mail: carma.ray.hogue@census.gov.

resources and reduce respondent burden by selecting a sample from U_{h1} with smaller sample size under the modified than under the traditional design. Let S_{hj} be a PPS sample of size n_{hj} from U_{hj}, $j = 1, 2$, $n_{h1} + n_{h2} = n_h$. Note that n_{h1} may still be larger than n_{h2} because N_{h1} is usually much larger than N_{h2}.

For unit $i \in U$, let y_i be a key survey variable (*e.g.*, the full-time employment, full-time payroll, part-time employment, part-time payroll, part-time hours), x_i be an auxiliary variable, say the same variable as y_i from the most recent census, and let z_i be the covariate used as the size variable in PPS sampling. The covariate values x_i and z_i are observed for all $i \in U$, whereas y_i is observed only for each sampled unit i.

The Horvitz-Thompson estimator of the unknown total $Y = \sum_{i \in U} y_i$ is

$$\hat{Y}_{\mathrm{HT}} = \sum_h \sum_j \sum_{i \in S_{hj}} y_i / \pi_i , \tag{1.1}$$

where π_i is the first-order inclusion probability of unit i in S_{hj}, a known function of z_i's. To utilize the auxiliary variable x_i and increase the accuracy of estimation of Y, the model-assisted approach (Särndal, Swensson and Wretman 1992) is adopted. Applying regression within each S_{hj} leads to the regression estimator of Y as

$$\hat{Y}_{\mathrm{reg},2} = \sum_h \sum_j \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right], \tag{1.2}$$

where $X_{hj} = \sum_{i \in U_{hj}} x_i$, $\hat{Y}_{hj} = \sum_{i \in S_{hj}} y_i / \pi_i$, $\hat{X}_{hj} = \sum_{i \in S_{hj}} x_i / \pi_i$, $\hat{N}_{hj} = \sum_{i \in S_{hj}} 1 / \pi_i$, and

$$\hat{\beta}_{hj} = \frac{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right) y_i / \pi_i}{\sum_{i \in S_{hj}} \left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2 / \pi_i} .$$

Alternatively, combining the two sub-strata S_{h1} and S_{h2} results in the following regression estimator. (A referee correctly points out that $\hat{Y}_{\mathrm{reg},1}$ in (1.3) is not the pooled estimator one would use if regression lines in stratum h were combined but the two sub-strata were not; however, it *is* the natural estimator when not only regression lines but also sub-strata are combined.)

$$\hat{Y}_{\mathrm{reg},1} = \sum_h \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right], \tag{1.3}$$

where $\hat{Y}_h = \sum_j \hat{Y}_{hj}$, $\hat{X}_h = \sum_j \hat{X}_{hj}$, $\hat{N}_h = \sum_j \hat{N}_{hj}$, and

$$\hat{\beta}_h = \frac{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right) y_i / \pi_i}{\sum_j \sum_{i \in S_{hj}} \left(x_i - \hat{X}_h / \hat{N}_h \right)^2 / \pi_i} .$$

Since both $\hat{Y}_{\mathrm{reg},1}$ and $\hat{Y}_{\mathrm{reg},2}$ are model-assisted estimators, they are consistent with respect to repeated sampling, whether or not the regression model holds. If the least-squares regression lines in two sub-strata

U_{hj}'s are the same, $\hat{Y}_{\text{reg},1}$ may be more efficient than $\hat{Y}_{\text{reg},2}$. On the other hand, if the regression lines are different, $\hat{Y}_{\text{reg},2}$ may be more efficient than $\hat{Y}_{\text{reg},1}$.

A decision-based method was proposed in Cheng *et al.* (2010), which applies hypothesis testing to decide whether we combine S_{h1} and S_{h2}. Within stratum h, the slopes of the regression lines in U_{h1} and U_{h2} are tested for equality. Let

$$\hat{\alpha}_{hj} = \frac{\hat{Y}_{hj} - \hat{\beta}_{hj}\hat{X}_{hj}}{\hat{N}_{hj}}, \quad \hat{\sigma}^2_{xe,hj} = \frac{n_{hj}}{\hat{N}^2_{hj}} \sum_{i \in S_{hj}} \left(x_i - \frac{\hat{X}_{hj}}{\hat{N}_{hj}} \right)^2 \frac{\left(y_i - \hat{\alpha}_{hj} - \hat{\beta}_{hj} x_i \right)^2}{\pi_i^2},$$

$$\hat{\sigma}^2_{xhj} = \sum_{i \in S_{hj}} \frac{\left(x_i - \hat{X}_{hj} / \hat{N}_{hj} \right)^2}{\pi_i \hat{N}_{hj}}, \quad t_h = \sqrt{n_h - 4}\left(\hat{\beta}_{h1} - \hat{\beta}_{h2} \right) \Bigg/ \sqrt{n_h \sum_{j=1}^{2} \frac{\hat{\sigma}^2_{xe,hj}}{n_{hj} \hat{\sigma}^4_{xhj}}}.$$

If $|t_h| > t_{1-\tau/2, n_h - 4}$, where $t_{1-\tau/2,\nu}$ is the $1 - \tau/2$ quantile of the t-distribution with ν degrees of freedom, then we reject the hypothesis of common slope and use $\hat{\beta}_{hj}$ (and set $\zeta_h = 1$). Here τ is a nominal significance level set by default to 0.05, although we will consider other choices of τ in the simulations section. The test-statistic definition involving $n_h - 4$ degrees of freedom is a slightly artificial choice designed to make the moderate-sample rejection probabilities closer to nominal, but the large-sample asymptotic distribution theory justifying this test is given in part (c) of Theorem 1. If $|t_h| \leq t_{1-\tau/2, n_h - 4}$, then we accept the hypothesis of common slope, combine sub-strata S_{h1} and S_{h2}, and use $\hat{\beta}_h$ (setting $\zeta_h = 0$). Tests are performed independently across strata $h = 1, \ldots, H$. The decision-based estimator of Y is then

$$\hat{Y}_{\text{dec}} = \sum_h \sum_j \zeta_h \left[\frac{N_{hj} \hat{Y}_{hj}}{\hat{N}_{hj}} + \hat{\beta}_{hj} \left(X_{hj} - \frac{N_{hj} \hat{X}_{hj}}{\hat{N}_{hj}} \right) \right] + \sum_h (1 - \zeta_h) \left[\frac{N_h \hat{Y}_h}{\hat{N}_h} + \hat{\beta}_h \left(X_h - \frac{N_h \hat{X}_h}{\hat{N}_h} \right) \right]. \quad (1.4)$$

Since two regression lines with a common slope can have different intercepts, one might test a further hypothesis regarding intercepts to decide whether to combine the two sub-strata. However, population points (x_i, y_i) falling on two parallel but not identical substratum regression lines would be discontinuous around the cut-off point between the two sub-strata U_{h1} and U_{h2}, which seems to occur only rarely in practical situations. In ASPEP, for example, Cheng *et al.* (2010) investigated the slopes and intercepts for sub-strata in 2002 and 2007 Census data sets, noting that the hypothesis of a common intercept could never be rejected when the hypothesis of a common slope could not be rejected. Thus, the decision-based estimator in (1.4) depends only on hypothesis testing for equality of sub-stratum regression slopes.

The two-stage estimators studied here are particular instances of procedures previously termed estimators following preliminary testing. There is a large literature on such procedures in surveys, including a bibliography by Bancroft and Han (1977), a book by Saleh (2006), and a treatment by Fuller (2009, Section 6.7). An idea from Saleh (2006) is to estimate coefficients by a convex combination of the estimated coefficients from the separate strata with proportions depending on a test statistic. Such smoothed estimators might be more efficient than our decision-based procedures. If the stratum-specific

intercepts and slopes were regarded as random, then a model-based empirical-Bayes approach to survey estimation might also be tried.

The decision-based estimators (1.4) are novel because they are model-assisted design-consistent in the survey-sampling context, making explicit use of the known substratum population sizes. In a somewhat similar spirit, Rao and Ramachandran (1974) previously made an exact comparison of the separate and combined ratio estimators under a ratio model similar to the regression model of this paper.

The purpose of this paper is to show some asymptotic and empirical properties of the estimators of Y described above and their variance estimators. Consistency and asymptotic normality of $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$ are established in Section 2, in terms of either design-based or model-assisted asymptotic theory. Although the first-order asymptotics favor $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{reg},1}$ may be better when some substratum sample sizes n_{h2} are moderate, a second-order asymptotic effect. The virtue of the decision-based estimator $\hat{Y}_{\text{dec}}$ is in adapting to be close to the better of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$. As the discussion in paragraph (III) of Section 4.4 indicates, simulations show that the benefit of this adaptivity is to reduce MSE up to a few percent under reasonable parameter settings, and by larger amounts in stranger settings.

Variance estimation for the decision-based estimator is treated in Section 3. While the asymptotic theory in Section 2 suggests that consistent variance estimators are obtained by substituting for unknown quantities in the asymptotic variance formulas, we also study bootstrap variance estimators suggested in Cheng *et al.* (2010), which are generally found to have better finite sample performance than the substitution estimators. Empirical results are presented in Section 4, with Section 4.4 providing interpretations and concluding remarks. All technical proofs are given in the Appendix.

2 Consistency and asymptotic normality

To consider asymptotics, we view the population U as one of a sequence of populations $\{U^{(m)}, m = 1, 2, \ldots\}$, where the number of units in $U^{(m)}$ increases to infinity as $m \to \infty$. This paper treats only the case of strata in which a large sample n_h is drawn; that is, we assume that for each stratum h, the sample size n_h depends on m and increases to infinity as $m \to \infty$, but we omit the index m for simplicity. All limiting processes are considered as $m \to \infty$. Following authors such as Isaki and Fuller (1982) and Deville and Särndal (1992), we term this a *superpopulation* asymptotic framework. Under the design-based framework considered in Section 2.1, the attribute vectors in the underlying populations need not be viewed as random vectors. However, under the model-assisted framework considered in Section 2.2, regression models are assumed for attribute vectors.

Since each estimator is a sum of independent estimators constructed within each stratum, for simplicity we present asymptotic results for the case of $H = 1$. The results and conclusions immediately apply to the case of a fixed H and can also be extended to the situation where H increases to infinity. (It is typical for large-scale surveys to have many strata, although the number of ASPEP government-by-type strata that were split into substrata was somewhat less than 100.) Since we only consider $H = 1$, we omit the index h for stratum in this section, *e.g.*, $n_{hj} = n_j$, $n_h = n$, $N_{hj} = N_j$, and $N_h = N$. Also, for

$j = 1, 2,$ the estimators $\hat{\beta}_j$ and $\hat{\beta}$ are defined by the displayed formulas following equations (1.2) and (1.3), with subscript h suppressed, together with

$$\hat{\mu}_{xj} = \hat{X}_j / \hat{N}_j, \quad \hat{\alpha}_j = \hat{Y}_j / \hat{N}_j - \hat{\beta}_j \hat{\mu}_{xj}, \quad \hat{\sigma}^2_{xj} = \hat{N}_j^{-1} \sum_{i \in S_j} \pi_i^{-1} \left(x_i - \hat{\mu}_{xj}\right)^2$$

$$\hat{\sigma}^2_{xe,j} = n_j \sum_{i \in S_j} \left(x_i - \hat{\mu}_{xj}\right)^2 \left(y_i - \hat{\alpha}_j - \hat{\beta}_j x_i\right)^2 \Big/ \left(\pi_i^2 \hat{N}_j^2\right).$$

Furthermore, for simplicity we consider asymptotics only under with-replacement sampling. The results can be applied to the case of without replacement sampling if the sampling fraction n/N is negligible.

2.1 Design-based asymptotic framework

First, we establish the asymptotic normality of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ under repeated sampling, that is, when y_i and x_i are fixed for $i \in U$, and S_j is a random PPS sample.

Theorem 1 Suppose that S_1 and S_2 are independent PPS samples with replacement from U_1 and U_2, respectively, where unit $i \in U_j$ has probability $p_{ij} = z_i \Big/ \sum_{i \in U_j} z_i > 0$ of being selected, and sampling weight $\pi_i^{-1} = 1/\left(n_j p_{ij}\right)$ for $j = 1, 2,$ and that the following four conditions hold, as the population sequence index m goes to ∞.

(C1) There exist constants φ_j and ω_j such that $\sqrt{n/n_j} \to \varphi_j$ and $N_j / N \to \omega_j$.

(C2) For $j = 1, 2,$ there exist constants μ_{yj}, μ_{xj} and β_j such that

$$\bar{Y}_j = Y_j / N_j = \sum_{i \in U_j} y_i / N_j \to \mu_{yj}, \bar{X}_j = X_j / N_j = \sum_{i \in U_j} x_i / N_j \to \mu_{xj}$$

exist, as do the limits $N_j^{-1} \sum_{i \in U_j} \left(x_i - \mu_{xj}\right)^2 \to \sigma^2_{xj} > 0,$ and in addition,

$$\left(\sqrt{n_j} / N_j\right) \sum_{i \in U_j} x_i \left(y_i - Y_j / N_j - \beta_j \left(x_i - X_j / N_j\right)\right) \to 0 \text{ as } n, N \to \infty.$$

(C3) The limits $D_{N_j} = \sum_{i \in U_j} p_{ij} b_{ij} b_{ij}^T / N_j^2 \to D_j$ exist, where for $i \in U_j$,

$$b_{ij} = \left[1/p_{ij} - N_j, x_i/p_{ij} - X_j, y_i/p_{ij} - Y_j\right]^T,$$

v^T denotes the vector transpose, and D_j is positive definite. The limit $\sigma^2_{xe,j} = \lim N_j^{-2} \sum_{i \in U_j} \left(x_i - \mu_{xj}\right)^2 \left(y_i - \alpha_j - \beta_j x_i\right)^2 / p_{ij}$ also exists, for $\alpha_j = \mu_{yj} - \beta_j \mu_{xj}$.

(C4) The elements of $\Lambda_j = \sum_{i \in U_j} p_{ij} c_{ij} c_{ij}^T / N_j^4$ form a bounded sequence, where for $i \in U_j$, $c_{ij} = \left[\left(1/p_{ij} - N_j\right)^2, \left(x_i/p_{ij} - X_j\right)^2, \left(y_i/p_{ij} - Y_j\right)^2\right]^T$.

Then, as $m \to \infty$, the following conclusions hold.

(a) For $j = 1, 2$, $\hat{\mu}_{xj} \to_P \mu_{xj}, \hat{\mu}_{yj} \to_P \mu_{yj}, \hat{\beta}_j \to_P \beta_j, \hat{\alpha}_j \to_P \alpha_j$, and $\hat{\sigma}^2_{xj} \to_P \sigma^2_{xj}$, where $\to_P$ denotes convergence in probability.

(b) The combined-stratum estimator $\hat{\beta}$ has the exact expression

$$\hat{\beta} = \frac{\sum_{j=1}^{2} \hat{\beta}_j \hat{\sigma}^2_{xj} \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)\left(\hat{Y}_2 - \hat{Y}_1\right) \hat{N}_1 \hat{N}_2 \big/ \left(\hat{N}_1 + \hat{N}_2\right)}{\sum_{j=1}^{2} \hat{\sigma}^2_{xj} \hat{N}_j + \left(\hat{X}_2 - \hat{X}_1\right)^2 \hat{N}_1 \hat{N}_2 \big/ \left(\hat{N}_1 + \hat{N}_2\right)} \tag{2.1}$$

and the in-probability limit

$$\beta = \frac{\sum_{j=1}^{2} \beta_j \sigma^2_{xj} \omega_j + (\mu_{x2} - \mu_{x1})\left(\mu_{y2} - \mu_{y1}\right) \omega_1 \omega_2}{\sum_{j=1}^{2} \sigma^2_{xj} \omega_j + (\mu_{x2} - \mu_{x1})^2 \omega_1 \omega_2}.$$

(c) $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right) \to_d N\left(0, \sigma^2_{xe,j} \big/ \sigma^4_{x,j}\right)$, where $\to_d$ denotes convergence in distribution, and $\hat{\sigma}^2_{xe,j} \to_P \sigma^2_{xe,j}$.

(d) For $k = 1, 2$,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},k} - Y\right) \big/ N \to_d N\left(0, \sigma^2_k\right) \tag{2.2}$$

where $\sigma^2_k = \sum_{j=1}^{2} a_{kj}^T D_j a_{kj}$ and

$$a_{1j} = \omega_j \varphi_j \left[-\left(\mu_y - \beta \mu_x\right), -\beta, 1\right]^T, \quad a_{2j} = \omega_j \varphi_j \left[-\left(\mu_{yj} - \beta_j \mu_{xj}\right), -\beta_j, 1\right]^T,$$

$\mu_x = \omega_1 \mu_{x1} + \omega_2 \mu_{x2}, \mu_y = \omega_1 \mu_{y1} + \omega_2 \mu_{y2}$, and D_j is given in condition (C3).

The conditions (C1)-(C4) of Theorem 1 provide a general formulation of the superpopulation framework for large-sample design-based statistical inference, within which the survey regression coefficients estimate well-defined frame-population descriptive parameters. The results in parts (a)-(b) show that the in-probability limits β_j, α_j of $\hat{\beta}_j, \hat{\alpha}_j$ have the standard interpretation as superpopulation least-squares slopes and intercepts. (These slope and intercept parameters also keep their usual model-based interpretations under the model (2.7) introduced in Section 2.2.) The asymptotic distribution theory for $\hat{\beta}_j$ in conclusion (c) allows us to deduce the large-sample behavior of $\hat{Y}_{\text{dec}}$ from that provided in (d) for $\hat{Y}_{\text{reg},k}$.

Under the further conditions

$$\beta_1 = \beta_2, \alpha_1 = \alpha_2, \tag{2.3}$$

it is clear from Theorem 1(b) that $\beta_j = \beta$, and $\sigma^2_1 = \sigma^2_2$ in (2.2), so that $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ are all asymptotically the same up to remainders of smaller order than $N/\sqrt{n}$, as we now show. Also, if

$\beta_1 \neq \beta_2$, then $\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}$ continues to be $o_P\left(N/\sqrt{n}\right)$, and the test of equality of slopes rejects, *i.e.*, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, and therefore $\hat{Y}_{\text{dec}}$ has the same asymptotic distribution as $\hat{Y}_{\text{reg},2}$, which is more efficient than $\hat{Y}_{\text{reg},1}$ according to the result in Section 2.2.

Theorem 2 Assume the same hypotheses (C1)-(C4) as in Theorem 1.

(a) When (2.3) holds, then as $m \to \infty$

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1\right) \to_d N\left(0, \sigma_d^2\right), \qquad \sigma_d^2 = \sum_{j=1}^{2} \frac{\sigma_{xe,j}^2}{\varphi_j^2 \sigma_{xj}^4}, \tag{2.4}$$

and the estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$ are all asymptotically normally distributed and equivalent in the sense that

$$\frac{n}{N^2}\left[\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)^2 + \left(\hat{Y}_{\text{reg},2} - \hat{Y}_{\text{dec}}\right)^2\right] \to_P 0. \tag{2.5}$$

(b) When $\beta_1 \neq \beta_2$, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$ and $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \to_d N\left(0, \sigma_2^2\right)$.

A more refined study of the asymptotic behavior of the estimators $\hat{Y}_{\text{dec}}$ can be undertaken in the spirit of Saleh (2006), as with contiguous or Pitman alternatives for non-survey statistical models, by assuming that $\sqrt{n}(\beta_1 - \beta_2) \to r$ for a constant r. Under this assumption, it can be shown that $\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2} = o_P\left(N/\sqrt{n}\right)$ and, therefore, the three centered and scaled estimators $\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)$, $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)$, and $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)$ all have the same asymptotic normal distribution with mean 0. Furthermore,

$$P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to \Phi\left(-z_{\tau/2} + r/\sigma_d\right) + \Phi\left(-z_{\tau/2} - r/\sigma_d\right), \tag{2.6}$$

where σ_d^2 is given in (2.4), and $z_{\tau/2}$ and Φ are respectively the standard normal percentage point and distribution function. Thus, $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$ has a limit different from 1. In particular, the limit in (2.6) equals τ when $\beta_1 = \beta_2$ (*i.e.*, when $r = 0$).

2.2 Model-assisted asymptotic setting

We elaborate in this section the behavior of estimators $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ under the assumed probabilistic model that the triples (x_i, y_i, z_i) in the finite population, $i \in U_j$, are independent and identically distributed (iid), where the size-variables $z_i > 0$ are used in defining PPS with-replacement draw probabilities $p_{ij} = z_i / \sum_{i' \in U_j} z_{i'}$, and where x_i and y_i follow the model

$$y_i = \alpha_j + \beta_j x_i + \varepsilon_i, \; i \in U_j, \tag{2.7}$$

with α_j and β_j as unknown intercept and slope parameters for the regression within stratum U_j. The errors $\varepsilon_i, i \in U_j$, are assumed to be iid with mean 0 and finite variance σ_ε^2 and to be independent of

(x_i, z_i), and the variables x_i for $i \in U_j$ are assumed to have finite variance. Also, to enable PPS sampling, we assume that $\max_{i \in U_j} n_j p_{ij} < 1$ with probability approaching 1 for large m, *i.e.*, for large n_j, N_j.

In this section, asymptotic properties of estimators $\hat{Y}_{\text{reg},k}, \hat{Y}_{\text{dec}}$ are considered with respect to the regression model and repeated sampling. By Theorem 1, the model-assisted estimators $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$ are still consistent and asymptotically normal for triples (x_i, y_i, z_i) iid within strata, since the conditions (C1)-(C4) are satisfied under moment assumptions on $z_i, 1/z_i$ even if model (2.7) is incorrect. However, the estimators $\hat{Y}_{\text{reg},k}$ are efficient when model (2.7) is correct.

Theorem 3 Assume model (2.7) along with (C1), with $E\left(x_i^4\right) < \infty, E\left(\varepsilon_i^4\right) < \infty, E\left(z_i\right) < \infty,$ and $E\left(\left(1 + x_i^4\right)/z_i^3\right) < \infty.$ Then all conclusions in Theorem 1 and Theorem 2 still hold. In particular, when $\beta_1 \neq \beta_2, \sigma_1^2,$ the asymptotic variance of $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N$, is larger than $\sigma_2^2,$ the asymptotic variance of $\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N$. Furthermore,

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)/N \rightarrow_d N\left(0, (1-\pi)\sigma_1^2 + \pi\sigma_2^2\right), \tag{2.8}$$

where π is the limit of $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right)$.

Note that π in (2.8) is equal to 1 when $\beta_1 \neq \beta_2$ and equal to τ when $\beta_1 = \beta_2$.

According to Theorem 3, under model (2.7), all three estimators defined in (1.2)-(1.4) have the same asymptotic efficiency when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ (condition (2.3)). Furthermore, $\hat{Y}_{\text{reg},1}$ is asymptotically worse than $\hat{Y}_{\text{reg},2}$ when $\beta_1 \neq \beta_2$. Thus, why would we not always use $\hat{Y}_{\text{reg},2}$?

The assertions in Theorem 3 are first-order asymptotic results. A more refined, second-order asymptotic result under the conditions in Theorem 3 and condition (2.3) when the sizes z_i are all equal is that, up to a term of order $n_1^{-2} + n_2^{-2}$,

$$\text{mse}\left(\frac{\hat{Y}_{\text{reg},1}}{N}\right) - \frac{\sigma_\varepsilon^2}{n} \leq \left[\text{mse}\left(\frac{\hat{Y}_{\text{reg},2}}{N}\right) - \frac{\sigma_\varepsilon^2}{n}\right]\left[1 - \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{nD_n}\right], \tag{2.9}$$

where mse is the mean squared error conditional on x_i's, $\bar{X}_j = N_j^{-1}\sum_{i \in U_j} x_i,$ and

$$D_n = \sum_{j=1}^{2}\sum_{i \in U_j}\left(x_i - \bar{X}_j\right)^2 + \frac{n_1 n_2 \left(\bar{X}_1 - \bar{X}_2\right)^2}{n}.$$

Result (2.9) indicates that, when weights are equal and $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2$, the finite sample performance of $\hat{Y}_{\text{reg},1}$ may be better than that of $\hat{Y}_{\text{reg},2}$ for moderate n_1 and n_2. See the simulation results in Section 4. The proof of (2.9) is a special case of a more general result in Slud (2012) and, thus, is omitted.

In applications, we do not know whether $\beta_1 = \beta_2$. Hence, the decision-based estimator $\hat{Y}_{\text{dec}}$ is an adaptive procedure to select a good estimator. In view of (2.8), the performance of $\hat{Y}_{\text{dec}}$ is close to (slightly worse than) that of $\hat{Y}_{\text{reg},2}$ when $\beta_1 \neq \beta_2$, and is close to (slightly worse than) that of $\hat{Y}_{\text{reg},1}$ when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. This is also supported by the simulation results in Section 4.

3 Variance estimation

It is common practice to report a variance estimate or standard error for each survey estimate. Variance estimation is also crucial for statistical inference when setting a confidence interval for an unknown parameter of interest.

The asymptotic results in Section 2 suggest a variance estimator for $\hat{Y}_{\text{reg},k}$ by substituting into (2.2) estimators for unknown quantities in σ_k^2. Since the total variance is a sum of H within-stratum variances, without loss of generality we consider one stratum $(H = 1)$. For $j = 1, 2,$ let

$$\hat{D}_{n_j} = \sum_{i \in S_j} \frac{\hat{b}_{ij}\hat{b}_{ij}^T}{(n_j - 1)\hat{N}_j}, \quad \hat{b}_{ij} = \left[1/p_{ij} - \hat{N}_j, x_i/p_{ij} - \hat{X}_j, y_i/p_{ij} - \hat{Y}_j\right]^T, \quad i \in S_j,$$

$$\hat{a}_{1j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y}_j - \hat{\beta}_j \bar{x}_j\right), -\hat{\beta}_j, 1\right]^T, \quad \hat{a}_{2j} = \frac{\hat{N}_j n^{1/2}}{\hat{N} n_j^{1/2}}\left[-\left(\bar{y} - \hat{\beta}\bar{x}\right), -\hat{\beta}, 1\right]^T,$$

$$\bar{y}_j = \hat{Y}_j/\hat{N}_j, \quad \bar{x}_j = \hat{X}_j/\hat{N}_j, \quad \bar{y} = \sum_{j=1}^{2}\hat{Y}_j/\left(\hat{N}_1 + \hat{N}_2\right), \quad \bar{x} = \sum_{j=1}^{2}\hat{X}_j/\left(\hat{N}_1 + \hat{N}_2\right).$$

Then, under the conditions in Theorem 1,

$$\hat{\sigma}_k^2 = \sum_{j=1}^{2}\hat{a}_{kj}^T\hat{D}_{n_j}\hat{a}_{kj} \to_p \sigma_k^2, \quad k = 1, 2.$$

That is, $\hat{\sigma}_k^2$ is consistent for σ_k^2. The results in Theorems 2 and 3 also show that $\hat{\sigma}_2^2$ is a consistent variance estimator for the decision-based estimator $\hat{Y}_{\text{dec}}$, because we have either $\sigma_1^2 = \sigma_2^2$ or $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$.

These substitution variance estimators, however, may not perform well when one of n_1 and n_2 is moderate (see Section 4). An alternative method is the bootstrap as suggested by Cheng *et al.* (2010). Let $\hat{\theta}$ be the estimator under consideration. Its bootstrap variance estimator can be obtained as follows.

1. Draw a bootstrap sample S_j^* as a simple random sample of size n_j with replacement from S_j, where S_1^* and S_2^* are independently obtained. If there are k_j self-representing units in S_j, as discussed in Section 4.1 below, then with-replacement samples of sizes $n_j - k_j$ are drawn, $j = 1, 2.$

2. The survey weights and observed data from the original data set are used to form a bootstrap data set $S_1^* \cup S_2^*$. From this dataset, calculate the bootstrap analog $\hat{\theta}^*$ of $\hat{\theta}$.
3. Independently repeat the previous steps B times to obtain $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$. The sample variance of $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ is the bootstrap variance estimator for $\hat{\theta}$.

Under the conditions in Theorems 1-2, the bootstrap variance estimators for $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ are consistent estimators. The proof for the bootstrap is similar to the proofs of the theorems and is omitted.

4 Simulation results for $H = 1$

Large sample theory as presented above is not adequate to tell whether the asymptotic results adequately describe the behavior of the estimators $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$ and their variance estimators in moderate samples, or whether $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{dec}}$ ever provide useful Mean-Squared-Error improvements in moderate sized samples. We present some simulation results to study these questions, as well as the small-sample issues arising in applying these methods in the context of the ASPEP survey.

In the simulations, values in the frame population U are either generated under some model or are taken from the 2002 and 2007 Government censuses with 2007 ASPEP sample weights. The first set of simulations (reported in Tables 4.1-4.6) summarizes average behavior over many model-generated frame populations. In the second set of artificial-data simulations, summarized in Table 4.8, the frame population remains fixed throughout the simulation. All frame populations consist of a single stratum $(H = 1)$ broken into two substrata $(j = 1, 2)$ according as a size variable falls below or above a specific quantile, usually the 0.8 quantile. Sampling from the frame populations is done PPS with-replacement in all simulations in this section.

4.1 Small sample considerations

Before proceeding to describe the simulations, we discuss some special features of PPS with-replacement (PPSWR) sampling which, when done in settings with small samples and unbalanced size variables, requires special computational handling. Numerically erratic results can arise when the small drawn samples are used stratumwise and then bootstrapped to estimate variances.

The weights $\pi_i^{-1} = 1/(n_j p_{ij})$ in PPSWR are all greater than 1 only when the single-draw probabilities $p_{ij} = z_i / \sum_{i' \in U_j} z_{i'}$ are all below $1/n_j$. To avoid anomalous small-sample results, and to maintain the relevance of PPSWR designs in imitating PPS without-replacement designs, any units $i \in U_j$ with $n_j p_{ij} \geq 1$ are made *self-representing* (SR), *i.e.*, are sampled with certainty but only once, and if there are k_j such units, then the probabilities $\{p_{ij} : i \in U_j, n_j p_{ij} < 1\}$ are renormalized to draw a size $n_j - k_j$ PPSWR sample. If any of the remaining renormalized probabilities are $\geq 1/(n_j - k_j)$, then their units also become self-representing and a new renormalization is done. This is repeated as often

as necessary. Thus, small samples with very unequal size-variable distributions may not be compatible with PPSWR sampling, a condition arising in some of the real-data ASPEP cases considered below.

Although a different choice could have been made, we conform with ASPEP practice in including all SR units in the fitting of the survey-weighted regression estimators $\hat{\beta}_2$ and $\hat{\beta}$. However, with this choice, PPSWR sampling followed by bootstrap resampling of small samples can lead to extremely erratic behavior, which must be recognized in summarizing the behavior of bootstrap variance estimators. The problem is that when a small number m of non-self-representing items are sampled PPSWR, in addition to a set of SR items, and then bootstrapped, the probability can be surprisingly large that there is only one unique non-SR item in the bootstrap sample, leading to very high bootstrap variability. This phenomenon was observed in the simulations reported below, with large-size substratum containing 20 or fewer elements and very skewed size-variables, either in the cases with lognormal or ASPEP x_i variables.

4.2 Artificial model-generated data

All of the artificial frame populations were generated with $N = 2{,}000$ iid triples (x_i, y_i, z_i) satisfying (2.7), for U_1 consisting of the $N_1 = 1{,}600$ for which x_i fell below their empirical 80'th percentile $c = \left(x_{(1{,}600)} + x_{(1{,}601)}\right)/2$, and U_2 consisting of the other 400 indices. In most cases, z_i were generated as $N(30 + x_i, 100)$ variates conditioned to be positive (which required occasional re-simulation in the lognormal-x_i models below) and were conditionally independent of y_i given x_i. (However, in some cases, unweighted samples were drawn by taking z_i identically equal.) PPS with-replacement stratified samples of sizes $(n_1, n_2) = (100, 50), (100, 20)$, or $(50, 20)$ were drawn in successive simulation runs, with size-variables z_i, from the same frame.

The models generating (x_i, y_i) are indexed as follows. In those with prefix **M1**, the predictors x_i are Gamma$(4, 0.1)$ distributed, with 0.8 quantile 55.2, while in the models **M2**, the x_i variables are Lognormal(1,6.25), with 0.8 quantile 22.3. The **M1** populations, and the **M2** models with suffix **E**, have conditional variance 100 for y_i given x_i, while the **M2** models without suffix **E** have conditional variance $20x_i$. Conditional means $E(y_i|x_i)$ are all linear, equal to $20 + 1.5x_i$ in models indexed **H0** and to $20 + x_i + 0.5(x_i - c)I_{[j=2]}$ within the substratum U_j in models **H1**. The intercepts of the regression models are so chosen that whether or not the slopes are the same, the lines intersect at $x = c$ (see the discussion in Section 1). Table 4.1 exhibits the average and standard deviation for the totals Y generated from the frame-population attributes $\{y_i\}_{i=1}^{2{,}000}$ under the various models. The variates x_i as well as the totals Y are much longer-tailed under the Lognormal models.

Table 4.1
Means and standard deviations for totals Y under simulation models.

	Gamma		Lognormal			
Model	**M1.H0**	**M1.H1**	**M2.H0**	**M2.H0E**	**M2.H1**	**M2.H1E**
E(Y)	160,000	123,177	225,603	225,603	173,485	173,485
SD(Y)	1,414.2	653.5	94,380	94,368	62,362	62,344

Simulated population models

M1.H0: $x_i \sim \text{Gamma}(4, 0.1)$ (shape parameter 4, scale 10),
$y_i \sim N(20 + 1.5x_i, 100)$ (variance 100), all $i \in U$.

M1.H1: $x_i \sim \text{Gamma}(4, 0.1)$, $y_i \sim N\left(20 + x_i + 0.5(x_i - c)\, I_{[x_i \geq c]}, 100\right)$, all i.

M2.H0: $\log(x_i) \sim N(1, 6.25)$, $y_i \sim N(20 + 1.5x_i, 20x_i)$, all i.

M2.H0E: $\log(x_i) \sim N(1, 6.25)$, $y_i \sim N(20 + 1.5x_i, 100)$, all i.

M2.H1: $\log(x_i) \sim N(1, 6.25)$, $y_i \sim N\left(20 + x_i + 0.5(x_i - c)\, I_{[x_i \geq c]}, 20x_i\right)$, all i.

M2.H1E: $\log(x_i) \sim N(1, 6.25)$, $y_i \sim N\left(20 + x_i + 0.5(x_i - c)\, I_{[x_i \geq c]}, 100\right)$, all i.

The simulation and bootstrap results in Tables 4.2-4.5 were generated by the following design and reporting scheme. For each population type, 60 distinct frame populations were generated, and 50 independent sampling experiments were conducted with each of those. In those cases where results of weighted and unweighted sampling were compared, these samples were drawn independently from the same set of 60 frame populations. Thus there were 3,000 independent replications for Monte Carlo averaging of statistical results, done for each of three different stratified sample sizes, and 400 bootstrap iterations were performed for each such generated sample.

Table 4.2
Empirical and estimated SD's and CI coverage, from model M1 simulations.

		M1.H0			M1.H1		
Sizes	**Stat**	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$	$\hat{Y}_{\text{reg},1}$	$\hat{Y}_{\text{reg},2}$	$\hat{Y}_{\text{dec}}$
100,50	SD_{MC}	1,785.5	1,794.3	1,788.0	1,817.6	1,773.5	1,774.4
	$\widehat{\text{SD}}_S$	1,757.1	1,751.5	1,755.6	1,794.6	1,735.2	1,735.8
	$\widehat{\text{SD}}_B$	1,752.4	1,762.0	1,758.4	1,788.1	1,742.9	1,747.0
	CP_S	94.47	94.37	94.50	93.93	93.73	93.77
	CP_B	94.60	94.53	94.67	93.93	94.03	94.07
100,20	SD_{MC}	1,930.0	1,944.8	1,934.0	2,008.4	1,944.4	1,960.4
	$\widehat{\text{SD}}_S$	1,888.3	1,876.6	1,884.1	1,944.4	1,861.0	1,866.5
	$\widehat{\text{SD}}_B$	1,878.8	1,901.4	1,895.8	1,936.1	1,885.6	1,897.9
	CP_S	94.20	93.83	94.13	93.53	93.20	93.07
	CP_B	93.80	94.00	93.97	93.60	93.83	93.97
50,20	SD_{MC}	2,583.5	2,610.7	2,593.5	2,591.3	2,522.8	2,535.4
	$\widehat{\text{SD}}_S$	2,509.2	2,490.8	2,505.1	2,562.2	2,465.0	2,474.5
	$\widehat{\text{SD}}_B$	2,498.5	2,538.0	2,522.9	2,550.3	2,508.5	2,525.6
	CP_S	93.70	93.13	93.57	93.97	93.63	93.43
	CP_B	93.63	93.73	93.87	93.83	93.77	94.10

Table 4.3
Empirical and estimated SD's and CI coverage, from model M2 simulations.

		M2.H0			M2.H1		
Sizes	**Stat**	$\hat{Y}_{reg,1}$	$\hat{Y}_{reg,2}$	$\hat{Y}_{dec}$	$\hat{Y}_{reg,1}$	$\hat{Y}_{reg,2}$	$\hat{Y}_{dec}$
100,50	SD_{MC}	3,400.1	3,475.4	3,406.8	3,481.9	3,483.8	3,482.2
	$\widehat{SD}_S$	3,420.6	3,400.0	3,417.0	3,537.8	3,405.0	3,463.7
	$\widehat{SD}_B$	3,590.0	3,715.2	3,623.4	3,852.0	3,921.9	3,898.4
	CP_S	95.10	93.43	94.83	95.03	93.40	94.13
	CP_B	95.67	95.77	95.77	95.63	95.77	95.70
100,20	SD_{MC}	5,655.2	6,184.0	5,698.6	5,853.0	6,181.1	5,955.6
	$\widehat{SD}_S$	5,644.9	5,575.7	5,640.9	5,798.3	5,587.3	5,697.3
	$\widehat{SD}_B$	5,565.1	6,687.3	5,857.8	5,907.8	6,838.0	6,466.6
	CP_S	93.83	88.47	93.40	92.77	88.30	90.70
	CP_B	92.33	93.67	93.37	92.63	94.33	94.17
50,20	SD_{MC}	5,773.2	6,319.2	5,833.9	5,934.2	6,230.6	6,009.8
	$\widehat{SD}_S$	5,800.2	5,677.2	5,785.8	6,012.6	5,755.4	5,919.2
	$\widehat{SD}_B$	5,728.5	6,825.2	6,086.0	6,102.2	6,978.1	6,522.1
	CP_S	94.60	88.67	93.97	94.07	89.37	92.27
	CP_B	93.40	94.23	94.27	93.47	95.03	94.80

Table 4.4
SD's for $\hat{Y}_{HT}$ *vs.* $\hat{Y}_{dec}$, and coverage for Bootstrap Percentile Confidence Intervals for $\hat{Y}_{dec}$, for $\tau = 0.05$ *vs.* 0.20, for models M1 and M2, H0 and H1.

		$\hat{Y}_{dec}, \tau = 0.05$		$\hat{Y}_{HT}$	$\hat{Y}_{dec}, \tau = 0.20$	
Model	**Samples**	SD_{MC}	CP_{BP}	SD_{HT}	SD_{MC}	CP_{BP}
M1.H0	100,50	1,788.0	94.23	2,774.0	1,745.5	94.60
	100,20	1,934.0	93.50	3,032.6	1,915.9	94.10
	50,20	2,593.5	93.17	3,000.7	2,500.1	94.43
M1.H1	100,50	1,774.4	93.70	2,387.3	1,737.3	94.43
	100,20	1,960.4	93.27	2,678.9	1,948.0	93.23
	50,20	2,535.4	93.90	3,035.0	2,509.8	94.23
M2.H0	100,50	3,406.8	95.20	4,160.0	3,398.8	94.83
	100,20	5,698.6	91.13	6,720.2	5,705.7	92.57
	50,20	5,833.9	92.60	7,080.0	5,979.8	92.17
M2.H1	100,50	3,482.2	95.13	4,393.6	3,423.9	94.03
	100,20	5,955.6	92.07	7,413.1	5,917.3	92.40
	50,20	6,009.8	92.33	7,840.4	6,105.6	92.17

Table 4.5
Comparisons of SD estimates and CI coverage for H0 and H1 for three lognormal settings, weighted (W) and unweighted (U) within M2, and weighted (E) within M2.E. CI % coverages are given for both the Bootstrap SD and Percentile Intervals.

Model	Size	Stat	SD	$\widehat{SD}_S$	$\widehat{SD}_B$	CP_S	CP_B	CP_{BP}
H0.W	100,50	$\hat{Y}_{reg,1}$	3,400.1	3,420.6	3,590.0	95.10	95.67	94.93
		$\hat{Y}_{reg,2}$	3,475.4	3,400.0	3,715.2	93.43	95.17	95.33
		$\hat{Y}_{dec}$	3,406.8	3,417.0	3,623.4	94.83	95.77	95.20
H0.U		$\hat{Y}_{reg,1}$	5,481.6	3,674.8	5,571.9	81.43	93.50	92.07
		$\hat{Y}_{reg,2}$	5,782.8	3,646.6	6,076.3	80.13	93.67	91.90
		$\hat{Y}_{dec}$	5,525.5	3,669.0	5,726.8	81.07	93.83	92.20
H0.E		$\hat{Y}_{reg,1}$	1,888.8	1,930.1	1,904.7	94.73	94.53	94.23
		$\hat{Y}_{reg,2}$	1,888.6	1,911.1	1,893.2	94.43	94.30	94.20
		$\hat{Y}_{dec}$	1,892.9	1,926.5	1,905.0	94.67	94.57	94.20
H0.W	50,20	$\hat{Y}_{reg,1}$	5,773.2	5,800.2	5,728.5	94.60	93.40	92.00
		$\hat{Y}_{reg,2}$	6,319.2	5,677.2	6,825.2	88.67	94.23	92.60
		$\hat{Y}_{dec}$	5,833.9	5,785.8	6,086.0	93.97	94.27	92.60
H0.U		$\hat{Y}_{reg,1}$	10,000.3	5,136.5	9,905.6	71.10	90.73	89.80
		$\hat{Y}_{reg,2}$	11,192.8	5,085.0	12,806.8	68.70	92.90	89.37
		$\hat{Y}_{dec}$	10,134.1	5,120.7	11,245.9	70.73	92.37	90.27
H0.E		$\hat{Y}_{reg,1}$	2,811.4	2,831.6	2,769.5	94.13	94.00	93.93
		$\hat{Y}_{reg,2}$	2,811.9	2,753.8	2,741.1	93.47	93.77	93.30
		$\hat{Y}_{dec}$	2,817.4	2,821.8	2,777.0	93.83	93.90	93.77
H1.W	100,50	$\hat{Y}_{reg,1}$	3,481.9	3,537.8	3,852.0	95.03	95.63	95.27
		$\hat{Y}_{reg,2}$	3,483.8	3,405.0	3,921.9	93.40	95.77	95.10
		$\hat{Y}_{dec}$	3,482.2	3,463.7	3,898.4	94.13	95.70	95.13
H1.U		$\hat{Y}_{reg,1}$	5,631.4	3,774.8	5,614.6	80.90	92.33	91.07
		$\hat{Y}_{reg,2}$	5,838.3	3,699.6	6,010.5	79.13	92.73	91.37
		$\hat{Y}_{dec}$	5,727.0	3,732.8	5,870.5	80.40	92.93	91.63
H1.E		$\hat{Y}_{reg,1}$	2,005.5	2,094.2	2,019.1	95.60	94.97	94.60
		$\hat{Y}_{reg,2}$	1,909.9	1,908.2	1,892.5	94.83	94.77	94.17
		$\hat{Y}_{dec}$	1,931.9	1,941.7	1,934.6	94.97	95.20	94.83
H1.W	50,20	$\hat{Y}_{reg,1}$	5,934.2	6,012.6	6,102.2	94.07	93.47	91.97
		$\hat{Y}_{reg,2}$	6,230.6	5,755.4	6,978.1	89.37	95.03	92.23
		$\hat{Y}_{dec}$	6,009.8	5,919.2	6,522.1	92.27	94.80	92.33
H1.U		$\hat{Y}_{reg,1}$	9,315.8	5,350.9	10,040.0	74.17	93.10	90.57
		$\hat{Y}_{reg,2}$	10,583.8	5,229.6	12,476.8	71.23	94.57	90.87
		$\hat{Y}_{dec}$	9,989.6	5,295.4	11,479.5	72.53	94.33	91.47
H1.E		$\hat{Y}_{reg,1}$	3,096.1	3,137.7	2,795.6	94.63	93.43	93.37
		$\hat{Y}_{reg,2}$	2,880.6	2,766.8	2,745.7	93.10	93.40	93.47
		$\hat{Y}_{dec}$	2,977.3	2,929.2	2,882.0	93.77	93.77	93.77

We calculated the following quantities for each combination of model, weighting, and sample size: the percentage biases of $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, $\hat{Y}_{\text{dec}}$ (with $\tau = 0.05$ in all tables except Table 4.4, and $\tau = 0.05$ or 0.20 in Table 4.4) as estimators of Y; the Monte Carlo standard deviations (SD), SD_{MC}, of these three estimators; the estimated SD's of the estimators, respectively using the substitution $(\widehat{\text{SD}}_S)$ and bootstrap $(\widehat{\text{SD}}_B)$ SD estimators described in Section 3; the coverage probability, CP_u, of the nominal 95% confidence intervals for $Y : \hat{Y} \pm 1.960 \cdot \widehat{\text{SD}}_u$, where $\hat{Y}$ is one of the three estimators of Y, and $u = S$ or B; and the bootstrap percentile confidence intervals (and their coverage percentages CP_{BP}) obtained from the empirical 0.025 and 0.975 quantiles of the (400) bootstrapped values of each of the three estimators $\hat{Y}$ of Y. In addition, we calculated empirical biases of the Horvitz-Thompson estimates $\hat{Y}_{\text{HT}}$ in (1.1) and their empirical standard deviations SD_{HT}. (Of these calculated quantities, only the biases are not shown, since all of the biases were well below 0.5% except in the model **M2.H1.U**, and even there the largest magnitude of bias was about 1%.) Two further statistics, computed and displayed in Table 4.6 for each of the estimators $\hat{Y}$ of Y, are the standard errors across randomly generated frame populations of the Monte Carlo and Bootstrap within-population estimated SD's of estimators $\hat{Y}$.

Table 4.6
Cross-population Standard errors of Empirical and Bootstrap SD's estimated for the estimators $\hat{Y}_{\text{reg},1}$, $\hat{Y}_{\text{reg},2}$, and $\hat{Y}_{\text{dec}}$, for selected models and weighting.

		$\hat{Y}_{\text{reg},1}$		$\hat{Y}_{\text{reg},2}$		$\hat{Y}_{\text{dec}}$	
Model	**Sizes**	**SD**	$\widehat{\text{SD}}_B$	**SD**	$\widehat{\text{SD}}_B$	**SD**	$\widehat{\text{SD}}_B$
M1.H0	100,50	198	35	196	35	197	35
	50,20	210	52	208	51	210	51
M1.H1	100,50	204	39	183	40	184	41
	50,20	319	57	298	62	302	62
M2.H0	100,50	404	345	450	383	405	351
	50,20	825	518	1,075	916	889	631
M2.H0.E	100,50	187	49	185	45	184	47
	50,20	294	85	293	71	298	82
M2.H1	100,50	409	409	410	421	408	414
	50,20	767	624	946	929	841	730
M2.H1.E	100,50	208	59	196	46	204	50
	50,20	258	141	261	82	239	102
M2.H1.U	100,50	1,676	1,351	1,773	1,539	1,726	1,467
	50,20	2,397	2,543	3,425	3,454	3,102	3,159

4.3 Real government-census data

Our simulations based on repeated sampling from real-data frames rely on a national state-wise dataset assembled by Yang Cheng. For the ASPEP survey of governments for sample year 2007, which was also a census year, the ASPEP frame is the same as the 2007 Census of Governments file. Our dataset consists of the 2002 and 2007 ASPEP variable values (full- and part-time employees, payroll and hours) derived from the censuses in those years, plus the 2007 sample weights and in-sample indicators for ASPEP. Weights equal to 1 imply that governmental units were self-representing (SR), in the sense that they were chosen

for inclusion with certainty in ASPEP. The size-variable z_i for PPS sampling within ASPEP is the sum of full- and part-time payroll from the most recent census, so we restrict attention to the 53,402 governmental units in the file for which this variable was positive. Table 4.7 gives the subcounty and special-district governmental types (the only ones that are subdivided into Small and Large unit substrata) in nine selected states, giving also the SR counts and numbers sampled in 2007. As mentioned in subsection 4.1, the final SR count for PPS with-replacement sampling can exceed the number of units initially chosen for certain inclusion, and the larger numbers, corresponding to the sample size actually drawn in 2007, are shown in the SR columns of Table 4.7. Inspection of this Table shows that several of the state by type combinations either have no population in a substratum or have too few non-SR units to be useful in simulating repeated samples. We take 15 as a rule-of-thumb minimum for the number of non-SR units, and suggest that substratum pairs with fewer non-SR units in the large-unit stratum should be collapsed without recourse to the decision-based strategy studied in this paper.

Table 4.7
Census population, ASPEP sample sizes and SR counts of Subcounty and Special-District governmental units by substratum in 2007, for 9 selected states.

	Subcounty					Special District				
	Small		Large			Small		Large		
	Pop	Samp	Pop	Samp	SR	Pop	Samp	Pop	Samp	SR
AL	378	15	55	45	26	0	0	400	102	64
CA	0	0	475	104	86	1,595	39	107	107	107
CO	0	0	265	34	18	627	16	65	55	33
FL	317	16	81	54	36	0	0	330	48	24
GA	461	17	49	36	20	0	0	293	70	32
MO	980	25	101	101	101	799	27	106	66	42
NY	1,473	25	69	69	69	606	16	33	23	4
PA	2,409	55	123	81	31	921	21	37	37	37
WI	1,702	36	129	71	44	281	16	61	40	20

For nine government-by-type combinations with 15 or more non-SR units and at least 17 non-sampled non-SR large-substratum units (except for AL, CO, and GA for which there were respectively 9, 10, and 11 non-sampled non-SR units), Table 4.8 displays results for the decision-based estimators and variance estimates in substratum pairs. In each of the state-type combinations, 3,000 stratified PPSWR samples of the indicated sizes were drawn from the ASPEP and government census frame described above, with x_i and y_i respectively the full-time payroll amount for the governmental unit as recorded in the 2002 and 2007 governmental censuses, and z_i the total (full-time plus part-time) payroll in 2002. Within each simulated sample, the estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ were calculated, and the empirical variances estimated. The variance of $\hat{Y}_{\text{dec}}$ was also estimated by the substitution formula and bootstrap methods as in the artificial-data simulations. (But note that, as described above, the bootstrap samples were drawn only from the non-SR units in each substratum sample.) The results are shown in Table 4.8. The relative efficiencies between the combined and separate stratified regression estimators can be gleaned from the corresponding ratio of SD's given in column 5 of the table. The remaining SD's shown are the empirical, substitution, and bootstrap SD estimators of $\hat{Y}_{\text{dec}}$.

Table 4.8
Summary of repeated-sampling simulations from ASPEP 2007 frame. Total full-time pay (Y) given in multiple of \$100 million, and estimated SD's of $\hat{Y}_{\text{dec}}$ given in columns 6-8 in units of \$1 million. SD_1/SD_2 in column 5 is ratio of empirical SD of $\hat{Y}_{\text{reg},1}$ over that of $\hat{Y}_{\text{reg},2}$.

State	Stratum	Y	Size	SD_1/SD_2	$\widehat{\text{SD}}$	$\widehat{\text{SD}}_S$	$\widehat{\text{SD}}_B$
AL	SubCty	1.2	25,46	2.14	4.90	3.67	5.71
CA	SpcDst	4.3	30,90	0.98	29.4	21.2	26.8
CO	SpcDst	0.6	25,55	1.14	3.77	2.58	3.00
FL	SubCty	4.3	25,54	1.16	11.9	9.4	12.2
GA	SubCty	1.5	25,38	1.15	4.38	3.26	4.88
MO	SpcDst	0.6	40,70	2.13	2.99	2.20	2.99
NY	SubCty	23.6	35,52	1.53	13.6	12.0	14.1
PA	SubCty	3.0	40,70	1.12	7.28	5.79	7.60
WI	SubCty	1.4	40,70	2.06	5.00	4.45	5.17

4.4 Discussion of simulation results

The following is a summary and interpretation of the results in the Tables, as well as of other results not shown.

(I) Many of the artificial-data simulations serve to confirm the large-sample theoretical results of the Theorems. It has already been mentioned that in Tables 4.2 and 4.3 the biases for all three Y-estimators $\left(\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}\right)$ are generally small. Within Table 4.2, referring to models with predictors and weights related to the Gamma distribution in models **M1**, the substitution and bootstrap variance estimators for each Y-estimator are quite accurate and close to one another, and the confidence intervals all have close to nominal coverage. Under both **M1**.**H0** and **M1**.**H1**, there is a tendency with smaller n_2 sample size for the $\widehat{\text{SD}}_S$ and $\widehat{\text{SD}}_B$ estimators to be slight underestimates of the actual or empirical SD's, but $\widehat{\text{SD}}_B$ seems to track SD more closely than $\widehat{\text{SD}}_S$ for $\hat{Y}_{\text{reg},2}$ and $\hat{Y}_{\text{dec}}$.

(II) The lognormal x_i values in models **M2** are much more dispersed and skewed than the values in **M1**, but the simulation results still support the asymptotic theory when $n_2 = 50$, although not when $n_2 = 20$. The substitution-estimator based confidence intervals for Y in terms of $\hat{Y}_{\text{reg},2}$ have coverage probability far too small when the substitution variance estimator is used. In Table 4.3, for each type of Y-estimator there is a pronounced tendency for the substitution variance estimator to underestimate the true (empirical) variance, and for the bootstrap estimator to overestimate.

Table 4.5 clarifies that the extreme behavior of variance estimators under models **M2** occurs partly because the predictors and y_i are dispersed and skewed, and partly because the size-variable used in PPS weighting shares these properties. The cases with suffix **W** in this Table are the same as in Table 4.3. The cases with suffix **E** have (x_i, z_i) the same as in Table 4.3, but the conditional variances of y_i given x_i have the constant value of 100; and with this change, the erratic behavior of SD estimators disappears. However, when the conditional y_i variances are as in the basic model **M2** but the PPSWR sampling is done *un*weighted, *i.e.*, with all z_i replaced by 1, the empirical and bootstrap SD estimators track each other and are very large, while the substitution variance estimator is too low by dramatic factors of $1/2$ to

3/4. This weird phenomenon applies equally to all three Y estimators. (However, an unweighted-sampling variant in model **M1** does not materially change the results from those shown in Table 4.2.)

(III) One objective of the simulations was to learn whether there can ever be any Mean-squared Error (MSE) benefit in using $\hat{Y}_{\text{reg},1}$ rather than $\hat{Y}_{\text{reg},2}$, without which there would be little motivation for $\hat{Y}_{\text{dec}}$. Indeed, the large-sample Theorems say that the main large-sample variance term is always optimal for $\hat{Y}_{\text{reg},2}$ (whether because it is the same as for $\hat{Y}_{\text{reg},1}$ under the null hypothesis or because it is strictly better under model (2.7) with distinct slopes). However, we indicated following Theorem 3, in the bound (2.9), that $\hat{Y}_{\text{reg},1}$ can have smaller second-order MSE than $\hat{Y}_{\text{reg},2}$, and the **H0** columns of Tables 4.2 and 4.3 do show a small but consistent SD advantage for $\hat{Y}_{\text{reg},1}$ *versus* $\hat{Y}_{\text{reg},2}$, an advantage which is more pronounced in **M2**. This advantage disappears under the fixed alternative **M1**.**H1** but interestingly, not under **M2**.**H1**. The slight but real conditional MSE advantage for $\hat{Y}_{\text{reg},1}$ when the substratum slopes are very close to equality is discussed further by Slud (2012).

The estimators $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}, \hat{Y}_{\text{dec}}$ considered here are of regression type, and it may be of interest to compare their MSE behavior in the simulated populations with that of the simpler Horvitz-Thompson estimator $\hat{Y}_{\text{HT}}$ in (1.1). All of these estimators are nearly unbiased, so that MSE's are essentially the same as variances, and a comparison of the third and fifth columns of Table 4.4 shows that the $\hat{Y}_{\text{HT}}$ variances are considerably larger than those of $\hat{Y}_{\text{dec}}$. The difference is least pronounced with the larger sample sizes, but even there is 30-55%. The advantage of $\hat{Y}_{\text{dec}}$ is still very pronounced in model **M2**, where model variances and distributional skewness are larger, but less so than in model **M1**.

(IV) The definition of $\hat{Y}_{\text{dec}}$ contains the arbitrary nominal significance level τ, which in all tables other than Table 4.4 was taken to be 0.05. As the large-sample theory suggests, the properties of the decision-based estimator fall between those of $\hat{Y}_{\text{reg},1}$ and $\hat{Y}_{\text{reg},2}$, and larger values of τ make $\hat{Y}_{\text{dec}}$ more often equal to $\hat{Y}_{\text{reg},1}$. As can be seen from comparison of columns 6 and 7 of Table 4.4, the choice τ = 0.20 seems in the simulated models to lead to very slightly smaller SD of $\hat{Y}_{\text{dec}}$ under model **M1**, but in model **M2** the SD is if anything larger at the smaller sample sizes. The conclusion is weak because the differences are quite small compared to the differences between SD's from one frame population to another. Our preference is to let smaller τ dictate the frequent pooling of substrata except when there are pronounced differences in estimated slope between the substrata. This finding that larger significance levels τ do not improve performance of $\hat{Y}_{\text{dec}}$ differs from the finding of Saleh (2006) that larger significance levels are highly beneficial in other preliminary-testing contexts.

(V) Table 4.6 gives information about the variability across frame populations of SD estimators for the Y estimators. The bootstrap variance estimators appear less susceptible to variation across frame populations, because the bootstrap averaging stabilizes them. The key finding in this table seems to be that the variability across frame populations is moderate except in the unweighted **M2** setting, where it is remarkably large. This seems to account for the extreme inflation of variances under **M2**.**U** seen in Table 4.5.

(VI) In many bootstrap applications with approximately normally distributed statistics, failure of coverage of normal-theory-based confidence intervals due to nonnormality of the bootstrapped statistic can be mitigated by using the bootstrap percentile (BP) intervals (Shao and Tu 1995, Section 4.1). In the

present simulations, Table 4.4 (columns 4 and 6) gives the coverage percentages of BP intervals for $\hat{Y}_{\text{dec}}$ in settings where Tables 4.2 and 4.3 give the coverages of normal-theory CI's based on the bootstrap-estimated SD. For whatever reason, the tables show that the normal-theory coverage CP_B tends systematically to be slightly below nominal and yet slightly larger than the BP interval coverage CP_{BP}. Thus, our simulations indicate the preference in this setting for the simpler interval $\hat{Y}_{\text{dec}} \pm 1.96 \cdot \widehat{\text{SD}}_B$.

(VII) It remains to draw lessons from the simulations with real government-census data in Section 4.3. The first necessary comment is that the spread and skewness of the full-time payroll predictors x_i and the total-payroll size-variable z_i are very large, much more like the lognormal models **M2** than the gamma models **M1**. Table 4.8 indicates (in column 5) a consistent MSE advantage for $\hat{Y}_{\text{reg},2}$ over $\hat{Y}_{\text{reg},1}$ except in the CA Special-district case, although the difference is small in the CO Special-district and the FL, GA and PA Subcounty cases. It is notable in almost all of these examples that the bootstrap SD estimator for $\hat{Y}_{\text{dec}}$ is more accurate than the substitution-formula estimator, despite the rather small numbers of sampled and unsampled non-SR units and (in several cases, as shown in Table 4.7) relatively large numbers of SR units. The substitution SD estimates are consistently too small while the bootstrap estimates are usually slightly high (*i.e.*, generally $\widehat{\text{SD}}_S < \widehat{\text{SD}} < \widehat{\text{SD}}_B$). The relative error of $\widehat{\text{SD}}_B$ *versus* $\widehat{\text{SD}}$ is no more than about 5% in these examples, except in the cases (AL, CO, GA) where there are particularly few non-sampled non-SR units in the large-unit substratum.

The large-unit substrata in ASPEP usually have small total frame population and often have relatively large numbers of SR units. While we have seen in these simulations that this does not quite invalidate inferences drawn with $\hat{Y}_{\text{reg},1}, \hat{Y}_{\text{reg},2}$ or $\hat{Y}_{\text{dec}}$, these statistics have distributions rather different from those of large-sample theory, and perhaps future substratum splits should allow slightly larger large-unit substrata for well-behaved statistical inferences.

More broadly, the simulation results indicate that the decision-based estimator with interval estimator defined from bootstrap variances is well-behaved and can be recommended except in extremely dispersed and skewed populations or in populations with large-unit sample sizes less than 20-25.

Acknowledgements

This paper describes research and analysis of the authors, and is released to inform interested parties and encourage discussion. Results and conclusions are the authors' and have not been endorsed by the Census Bureau. We would like to thank three referees and an associate editor for helpful comments and suggestions which improved the paper. Jun Shao's research was partially supported by the NSF Grant DMS-1007454.

Appendix

Proof of Theorem 1. Under PPS sampling, $\pi_i = n_j p_{ij}$ for unit $i \in U_j$, and on each with-replacement draw, the sampled index $i_t \in U_j, t = 1, \ldots, n_j$ has $P(i_t = i) = p_{ij}$ for each $i \in U_j$. By calculating

the means and variances (under repeated sampling) of $\hat{N}_j$, $\hat{X}_j$, $\hat{Y}_j$, $N_j^{-1}\sum_{i\in S_j} x_i y_i/\pi_i$ and $N_j^{-1}\sum_{i\in S_j} x_i^2/\pi_i$, we find the variances to be of order n_j^{-1} by means of the limits in (C2)-(C3) and the bounds in (C4). The assertions in part (a) follow directly.

For assertion (b), we have by definition of $\hat{\beta}$ that

$$\hat{\beta} = \frac{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right) y_i/\pi_i}{\sum_{j=1}^{2}\sum_{i\in S_j}\left(x_i - \hat{\mu}_{x,j} + \hat{\mu}_{x,j} - \left(\hat{X}_1 + \hat{X}_2\right)/\left(\hat{N}_1 + \hat{N}_2\right)\right)^2/\pi_i}$$

$$= \frac{N^{-1}\left(\sum_{j=1}^{2}\hat{\beta}_j\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)\left(\hat{\mu}_{x1} - \hat{\mu}_{x2}\right)\left(\hat{\mu}_{y1} - \hat{\mu}_{y2}\right)\right)}{N^{-1}\left(\sum_{j=1}^{2}\hat{\sigma}_{xj}^2\hat{N}_j + \left(\hat{N}_1\hat{N}_2/\left(\hat{N}_1 + \hat{N}_2\right)\right)\left(\hat{\mu}_{x1} - \hat{\mu}_{x2}\right)^2\right)},$$

from which the equality (2.1) in (b) follows immediately by substituting the limits in part (a) along with the limits $N_j/N \to \omega_j$.

Let Σ_N be the block diagonal matrix with two diagonal blocks D_{N_1} and D_{N_2}, and for $j = 1, 2$, let

$$\begin{aligned}
\Omega_{1j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{1}{p_{ij}} - N_j\right),\ \Omega_{2j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{x_i}{p_{ij}} - X_j\right),\\
\Omega_{3j} &= \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\left(\frac{y_i}{p_{ij}} - Y_j\right),\ \Omega_{4j} = \frac{1}{N_j\sqrt{n_j}}\sum_{i\in S_j}\frac{x_i - \mu_{x,j}}{p_{ij}}\left(y_i - \alpha_j - \beta_j x_i\right).
\end{aligned} \tag{A.1}$$

Since S_1 and S_2 are independent, $\{\Omega_{k1}\}_{k=1}^4$ is independent of $\{\Omega_{k2}\}_{k=1}^4$. Note that, here and throughout this proof, sums over $i \in S_j$ used to define $\hat{X}_j$, $\hat{Y}_j$, Ω_{kj}, and variance estimators should be understood as sums *with multiplicity* in view of the with-replacement PPS sampling framework. Condition (C4) makes Liapounov's Central Limit Theorem applicable to show that

$$\Sigma_N^{-1/2}\left[\Omega_{11}, \Omega_{21}, \Omega_{31}, \Omega_{12}, \Omega_{22}, \Omega_{32}\right]^T \to_d N(0, I_6),\ \Omega_{4j} \to_d N\left(0, \sigma_{xe,j}^2\right), \tag{A.2}$$

where I_6 is the 6×6 identity matrix, and $\sigma_{xe,j}^2$ is given in the statement of (d). The limits defining the asymptotic variances in (A.2) exist according to (C3).

Proof of (c). It is straightforward to check from the definition that

$$\begin{pmatrix}\hat{\beta}_j - \beta_j\\ \hat{\alpha}_j - \alpha_j\end{pmatrix} = \frac{1}{\hat{N}_j\hat{\sigma}_{xj}^2}\sum_{i\in S_j}\begin{pmatrix}x_i - \hat{\mu}_{xj}\\ \hat{\sigma}_{xj}^2 - (x_i - \hat{\mu}_{xj})\hat{\mu}_{xj}\end{pmatrix}\frac{y_i - \alpha_j - \beta_j x_i}{\pi_i}.$$

Since it was established in (a) that $\hat{\sigma}_{xj}^2 \to_P \sigma_{xj}^2$ and $\hat{N}_j/N_j \to_P 1$, it follows that the limiting distribution of $\sqrt{n_j}\left(\hat{\beta}_j - \beta_j\right)$ is the same as that of

$$\sqrt{n}_j \left(N_j \sigma_{xj}^2\right)^{-1} \sum_{i \in S_j} \left(x_i - \mu_{xj}\right)\left(y_i - \alpha_j - \beta_j x_i\right)/\pi_i ,$$

which is clearly the same as that of $\sigma_{xj}^{-2}\Omega_{4j}$ in (A.1). The first assertion of (c) follows immediately from (A.2). The consistency of $\hat{\sigma}_{xe,j}^2$ follows by noting by (a) that

$$\hat{\sigma}_{xe,j}^2 - N_j^{-2} \sum_{i \in S_j} \frac{\left(x_i - \mu_{xj}\right)^2}{\pi_i p_{ij}} \left(y_i - \alpha_j - \beta_j x_i\right)^2 \to_P 0. \tag{A.3}$$

The second term on the left-hand side of (A.3) has PPS with-replacement sampling variance calculated to be bounded by $1/n_j$ according to (C4), and by (C3) has expectation converging to $\sigma_{xe,j}^2$.

Proof of (d). From (1.2) and (a), $\left(\hat{Y}_{\text{reg},2} - Y\right)/N \to_P 0,$ which can also be seen from the representation

$$\begin{aligned}
\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N &= \frac{\sqrt{n}}{N} \sum_{j=1}^{2} \left[\frac{N_j \hat{Y}_j}{\hat{N}_j} - Y_j + \hat{\beta}_j \left(X_j - \frac{N_j \hat{X}_j}{\hat{N}_j} \right) \right] \\
&= \frac{\sqrt{n} N_1^2}{\sqrt{n_1} N \hat{N}_1} \left[\left(-\bar{Y}_1 + \hat{\beta}_1 \bar{X}_1\right)\Omega_{11} - \hat{\beta}_1 \Omega_{21} + \Omega_{31} \right] \\
&\quad + \frac{\sqrt{n} N_2^2}{\sqrt{n_1} N \hat{N}_2} \left[\left(-\bar{Y}_2 + \hat{\beta}_2 \bar{X}_2\right)\Omega_{12} - \hat{\beta}_2 \Omega_{22} + \Omega_{32} \right] \\
&= d_{n1}^T \bar{\Omega}_1 + d_{n2}^T \bar{\Omega}_2 ,
\end{aligned}$$

where the second equality follows from the notational definitions of Ω_{kj} along with $\pi_i = n_j p_{ij}$, $\hat{Y}_j = \sum_{i \in S_j} y_i/\pi_i$, $\hat{X}_j = \sum_{i \in S_j} x_i/\pi_i$, and the third from

$$d_{nj} = \frac{\sqrt{n} N_j^2}{\sqrt{n_j} N \hat{N}_j} \left[-\bar{Y}_j + \hat{\beta}_j \bar{X}_j, -\hat{\beta}_j, 1 \right]^T, \quad \bar{\Omega}_1 = \left[\Omega_{11}, \Omega_{21}, \Omega_{31}\right]^T, \quad \bar{\Omega}_2 = \left[\Omega_{21}, \Omega_{22}, \Omega_{32}\right]^T .$$

By (A.2), $\bar{\Omega}_1 = O_p(1)$ and $\bar{\Omega}_2 = O_p(1)$. By condition (C2), $d_{nj}^T = a_{2j}^T + o_p(1)$. Therefore, by (A.2), condition (C3) and the delta method,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1) \to_d N\left(0, \sigma_2^2\right),$$

where the asymptotic variance $\sigma_2^2 = \sum_{j=1}^{2} a_{2j}^T D_j a_{2j}$ is consistently estimated by

$$\frac{n}{N^2} \sum_{j=1}^{2} \sum_{i \in S_j} \frac{1}{\pi_i^2} \left(y_i - \hat{\beta}_j x_i - \left(\hat{Y}_j - \hat{\beta}_j \hat{X}_j\right)/\hat{N}_j\right)^2 ,$$

which agrees with formula (9) of Cheng *et al.* (2010). The proof that $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)/N \to_d N\left(0, \sigma_1^2\right)$ is similar.

Proof of Theorem 2. By Theorem 1 conclusion (c),

$$\sqrt{n}\left(\hat{\beta}_2 - \hat{\beta}_1 - \beta_2 + \beta_1\right) \to_d N\left(0, \sum_{j=1}^{2} \sigma_{xe,j}^2 \big/ \left(\varphi_j^2 \sigma_{xj}^4\right)\right). \tag{A.4}$$

The conclusion (2.4) in (a) of this Theorem follows immediately.

In the proof of Theorem 1, we showed that

$$\sqrt{n}\left(\hat{Y}_{\text{reg},2} - Y\right)\big/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.5}$$

where the constant vectors a_{kj} (and μ_x, μ_y) were defined in Theorem 1 (d). Similarly,

$$\sqrt{n}\left(\hat{Y}_{\text{reg},1} - Y\right)\big/N = a_{11}^T \bar{\Omega}_1 + a_{12}^T \bar{\Omega}_2 + o_p(1). \tag{A.6}$$

When (2.3) holds, $\beta_j = \beta$ (by Theorem 1 (b)) and $\mu_y - \beta\mu_x = \sum_{j=1}^{2} \omega_j \left(\mu_{yj} - \beta_j \mu_{xj}\right) = \mu_{y2} - \beta\mu_{x2}$, so that $a_{1j} = a_{2j}$ for $j = 1, 2$. It follows immediately from (A.5)-(A.6) that $\sqrt{n}\left(\hat{Y}_{\text{reg},1} - \hat{Y}_{\text{reg},2}\right)\big/N \to_P 0$, and therefore that the estimators $\hat{Y}_{\text{reg},k}$ have the same asymptotic distribution, which was shown to be normal in Theorem 1 (d). Finally, the definition of $\hat{Y}_{\text{dec}}$ implies that $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},1} \text{ or } \hat{Y}_{\text{reg},2}\right) = 1$ and (A.5)-(A.6) imply

$$\sqrt{n}\left(\hat{Y}_{\text{dec}} - Y\right)\big/N = a_{21}^T \bar{\Omega}_1 + a_{22}^T \bar{\Omega}_2 + o_p(1), \tag{A.7}$$

which completes the proof of (2.5) in (a).

Proof of (b). If $\beta_1 \neq \beta_2$, then (A.4) implies that $P\left(\hat{Y}_{\text{dec}} = \hat{Y}_{\text{reg},2}\right) \to 1$, *i.e.*, that the t-test for equality of $\hat{\beta}_j$ rejects with certainty in the limit. Then (A.7) continues to hold, and the asymptotic distribution of $\hat{Y}_{\text{dec}}$ is still as same as that of $\hat{Y}_{\text{reg},2}$.

Proof of Theorem 3. In this Theorem, the hypotheses (C2)-(C4) are replaced by the assumptions that the iid triples (y_i, x_i, z_i) satisfy moment conditions and the model (2.7). The assertions in (C2)-(C4) are then results holding with probability tending to 1 with large n, N which are established with the aid of the (strong) law of large numbers.

Beyond the conclusions of Theorems 1-2, it remains to show that $\hat{Y}_{\text{reg},2}$ has a smaller asymptotic variance than $\hat{Y}_{\text{reg},1}$. Let $\vartheta = (\vartheta_1, \vartheta_2)$ and

$$F_j(\vartheta) = [-\vartheta_1, -\vartheta_2, 1] D_j [-\vartheta_1, -\vartheta_2, 1]^T.$$

According to the definition of σ_1^2 and σ_2^2 in (2.2), it suffices to show that $F_j(\vartheta)$ has its minimum value at $\vartheta = (\alpha_j, \beta_j)$. We now prove this for $j = 1$. The proof for $j = 2$ is similar. Let $m_{ii'}$ be the (i, i') element of D_1. Since D_1 is symmetric and positive definite under condition (C3), $m_{12} = m_{21}$ and there

exists a unique $\theta^* = (\theta_1^*, \theta_2^*)$ such that $F_1(\theta^*) = \min_\vartheta F_1(\vartheta)$ and $\partial F_1(\vartheta)/\partial\vartheta^T|_{\vartheta=\theta^*} = 0$. This implies that θ^* is the solution to the following two equations:

$$m_{11}\vartheta_1 + m_{12}\vartheta_2 = m_{13}, \quad m_{12}\vartheta_1 + m_{22}\vartheta_2 = m_{23} \tag{A.8}$$

Therefore, it suffices to show that $\theta^* = (\alpha_1, \beta_1)$. Since D_1 is positive definite, the equation system (A.8) has a unique solution. By the definition of D_1,

$$\begin{aligned} m_{11}\alpha_1 + m_{12}\beta_1 &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)^2 p_{i1}\alpha_1 + \sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{x_i}{p_{i1}} - X_1\right)p_{i1}\beta_1\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

and

$$\begin{aligned} m_{13} &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)\left(\frac{y_i}{p_{i1}} - Y_1\right)p_{i1}\right] \\ &= \lim_{N_1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 + \beta_1 x_i + \varepsilon_i - N_1\alpha_1 p_{i1} - \beta_1 p_{i1} X_1)\right] \\ &= \lim_{N1\to\infty} \frac{1}{N_1^2}\left[\sum_{i\in U_1}\left(\frac{1}{p_{i1}} - N_1\right)(\alpha_1 - N_1\alpha_1 p_{i1} + \beta_1 x_i - \beta_1 p_{i1} X_1)\right], \end{aligned}$$

where the last equality follows from the assumption that ε_i is independent of x_i and z_i and has mean 0 and a finite variance, and each of the sequences z_i, $1/z_i$, and x_i/z_i is iid with finite expectation. Therefore, $m_{11}\alpha_1 + m_{12}\beta_1 = m_{13}$. Similarly one proves that $m_{12}\alpha_1 + m_{22}\beta_2 = m_{23}$. Therefore, (α_1, β_1) is the unique solution to equation system (A.8), *i.e.*, $F_1(\vartheta)$ achieves its minimum value at $\vartheta = (\alpha_1, \beta_1)$. Hence, $\sigma_2^2 < \sigma_1^2$. This finishes the proof of Theorem 3.

References

Bancroft, T., and Han, C.-P. (1977). Inference based on conditional specifications: A note and a bibliography. *International Statistical Review*, 45, 117-127.

Cheng, Y., Corcoran, C., Barth, J. and Hogue, C. (2009). An estimation procedure for the new public employment survey design. Washington, DC: American Statistical Association. *Survey Research Methods Section*, American Statistical Association, 3032-3046.

Cheng, Y., Slud, E. and Hogue, C. (2010). Variance estimation for decision-based estimators with application to the annual survey of public employment and payroll. *Government Statistics Section of the American Statistical Association*. Vancouver: American Statistical Association.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A. (2009). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Isaki, C., and Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Rao, J.N.K., and Ramachandran, V. (1974). Comparison of the separate and combined ratio estimators. *Sankhyā*, C, 36, 151-156.

Saleh, A.K. Md. (2006). *Theory of Preliminary Test and Stein-type Estimation, with Applications*. Hoboken: Wiley-Interscience.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., and Tu, D. (1995) *The Jackknife and Bootstrap.* New York: Springer.

Slud, E.V. (2012). Moderate-sample behavior of adaptively pooled stratified regression estimators. U.S. Census Bureau preprint.